PATHODC-VQA: CONSTRUCTING DIFFICULTY-CONTROLLED PATHOLOGY VQA DATA FROM SPARSE CAPTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Pathology vision–language models (VLMs) require domain-specific supervision for diagnostic reasoning, yet many public resources contain only information-sparse image–caption pairs. Naively converting these pairs into visual question answering (VQA) often produces caption paraphrases with little control over diagnostic difficulty. We introduce **PathoDC**, a three-stage pipeline that converts diagnostically eligible pairs into difficulty-controlled supervision. PathoDC first constructs structured Verified Diagnostic Knowledge (VDK) through automated generation and auditing. It then uses a VLM probe to organize candidates into single-image-answerable *Easy* cases and two evidence-sensitive hard tracks: knowledge-conditioned *Hard-K* and auxiliary-image-contrastive *Hard-V*. Finally, it generates track-conditioned rationales and applies an Easy→Hard-K→Hard-V training curriculum. This pipeline produces **PathoDC-VDK** with **6.3k** records and **PathoDC-VQA** with **109k** samples. Supervised fine-tuning of Qwen2.5-VL Instruct yields **PathoDC-3B** and **PathoDC-7B**. Using only about one-fifth of Patho-R1's SFT data and no reinforcement learning, PathoDC achieves competitive performance on public pathology benchmarks. At the 7B scale, PathoDC-7B exceeds Patho-R1-7B on Quilt-VQA and performs comparably on PathVQA, while PathoDC-3B surpasses Patho-R1-3B on PathVQA. On the case-disjoint **PathoDC-Bench**, the two models improve their backbone weighted-average accuracy by up to **42.34** points and outperform the corresponding Patho-R1 models by **9.89** and **21.17** points, respectively. Although the gains are not uniform across all public VQA and PathMMU evaluations, the results show that evidence-aware difficulty control can improve pathology VLM supervision, with the strongest gains on diagnostic evaluations aligned with the proposed evidence taxonomy.

1 INTRODUCTION

Training vision–language models (VLMs) (Chen et al., 2024b; Bai et al., 2025) for pathology diagnostic reasoning requires supervision that extends beyond short, descriptive figure captions. A capable diagnostic model must combine visual perception, disease knowledge, and contrastive disambiguation. Large-scale public image–caption corpora, such as PathCap and PathGen (Sun et al., 2024; 2025c), provide broad visual coverage, but they were curated primarily for contrastive pretraining or image captioning rather than diagnostic reasoning. Their captions are therefore often information-sparse—emphasizing staining protocols, magnifications, or experimental context—and lack an explicit, structured disease label field.

Repurposing these corpora for visual question answering (VQA) exposes two supervision gaps. **First, a knowledge deficit:** without explicit diagnostic priors, generated answers and rationales can become caption paraphrases or unsupported medical claims. **Second, a supervision-granularity gap:** a single VQA format obscures distinct evidence dependencies. We operationalize candidates as single-image-answerable Easy cases and two evidence-sensitive hard tracks: Hard-K, whose probe performance benefits from explicit disease knowledge, and Hard-V, constructed with auxiliary-image contrast to expose visually confusable alternatives. We treat the two hard tracks as complementary evidence axes rather than a scalar difficulty ladder, preserving the diagnostic skill targeted by each example.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

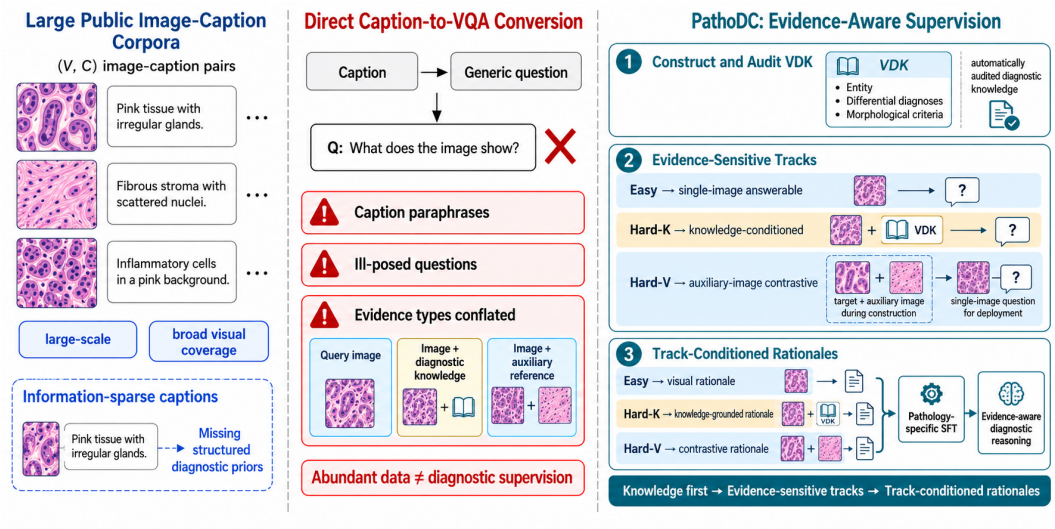


Figure 1: Motivation and overview of PathoDC. Public image–caption corpora provide broad visual coverage but sparse diagnostic supervision, and direct VQA conversion often conflates distinct evidence types. PathoDC instead constructs automatically audited diagnostic knowledge, organizes candidates into Easy, Hard-K, and Hard-V tracks, and generates track-conditioned rationales for SFT. Auxiliary images in Hard-V are used during data construction, while the resulting questions retain a single-image deployment format.

To address these gaps, we introduce **PathoDC**, which constructs supervision in layers: *knowledge first, evidence-sensitive difficulty second, and track-conditioned rationales third* (Fig. 1). Its central insight is that diagnostic difficulty should be organized not only by apparent hardness but also by the type of evidence targeted during construction. **Stage I** upgrades eligible image–caption pairs into Verified Diagnostic Knowledge (VDK) through automated generation and auditing: a context-aware store reuses disease-level rules, text–visual audits filter unsupported claims, and delexicalized serialization reduces direct answer copying. **Stage II** uses a probe VLM to operationally partition candidates into Easy and Hard cases. Hard-K items are conditioned on VDK, while Hard-V items use Type-A (intra-class) and Type-B (inter-class mimic) neighbors during construction while retaining single-image wording for deployment. **Stage III** attaches track-conditioned rationales for supervised fine-tuning (SFT). This design aligns the form of supervision with the evidence type that each track is intended to emphasize.

Our main contributions are threefold:

- We propose a three-stage pathology VQA pipeline—Verified Diagnostic Knowledge (VDK), an Easy/Hard split whose Hard items are constructed from knowledge and auxiliary-image axes, and axis-conditioned SFT rationales—that upgrades sparse captions into structured diagnostic supervision without manual textbook curation.

- We release the **PathoDC** suite: **PathoDC-VDK** (**6.3k** automatically audited knowledge records), **PathoDC-VQA** (**109,767** Easy/Hard-K/Hard-V SFT samples), and **PathoDC-Bench**, a case-disjoint Easy/Hard diagnostic benchmark designed around the same taxonomy.

- Fine-tuning Qwen2.5-VL Instruct on PathoDC-VQA yields **PathoDC-3B** and **PathoDC-7B**. Using approximately one-fifth of Patho-R1's SFT data and no reinforcement learning, PathoDC-7B exceeds Patho-R1-7B on Quilt-VQA and performs comparably on PathVQA, while PathoDC-3B surpasses Patho-R1-3B on PathVQA. On PathoDC-Bench, both models further outperform their corresponding Patho-R1 baselines; stage and component ablations characterize the effects of evidence type, rationale design, and curriculum ordering.

2 Related Work

Medical multimodal large language models. Recent medical MLLMs adapt general vision–language architectures to clinical imagery and text (Xiao et al., 2025). Early systems such as LLaVA-Med and Med-Flamingo demonstrate that instruction tuning or few-shot multimodal prompting can transfer biomedical knowledge from paired corpora (Li et al., 2023; Moor et al., 2023). Generalist models further scale this paradigm: BiomedGPT unifies diverse biomedical modalities under one foundation model (Zhang et al., 2024), while HuatuoGPT-Vision injects large-scale medical visual knowledge into multimodal LLMs (Chen et al., 2024a). More recent efforts emphasize open medical VLMs and explicit reasoning, including MedGemma and reinforcement-learning–based MedVLM-R1 (Sellergren et al., 2025; Pan et al., 2025). Despite rapid progress, these models are typically optimized for broad clinical coverage rather than pathology-specific diagnostic workflows; they rarely convert sparse image–caption supervision into an epistemic curriculum that separates perception, knowledge-conditioned deduction, and contrastive disambiguation.

Pathology multimodal large models. Computational pathology has developed both contrastive foundation encoders and generative assistants. CONCH provides a strong vision–language representation for histopathology (Lu et al., 2024a), and PathChat shows that a multimodal generative copilot can support interactive pathology workflows (Lu et al., 2024b). On the data and instruction side, PathAsst, Quilt-LLaVA, and PathGen construct large pathology image–text resources and generative assistants (Sun et al., 2024; Seyfioglu et al., 2024; Sun et al., 2025c), while CPath-Omni and PathOrchestra push toward unified multimodal models spanning patch- and slide-level tasks (Sun et al., 2025a; Yan et al., 2025). Reasoning-centric pathology VLMs such as Patho-R1 and PathVLM-R1 further improve diagnostic answers via reinforcement learning (Zhang et al., 2026; Wu et al., 2025). However, most prior work either emphasizes large-scale pre-training/instruction corpora or post-hoc RL, and still under-addresses how to upgrade information-sparse public captions into *verified* diagnostic priors with difficulty-controlled, track-conditioned rationales.

Curriculum learning and rationale supervision. Curriculum methods schedule examples by difficulty rather than mixing all skills in one pool (Bengio et al., 2009). Chain-of-thought prompting and STaR-style bootstrapping attach intermediate reasoning to answers (Wei et al., 2022; Zelikman et al., 2022), and VisCoR distills dual-image contrast into a single-image rationale (Pan et al., 2026). These techniques typically assume already well-posed items; they do not specify how to recover a verified diagnostic prior from sparse captions, nor how to admit Easy items and two unordered Hard construction axes from operational probe criteria before writing rationales. PathoDC applies them only after VDK construction and epistemic partitioning, which is the supervision gap this paper targets.

3 Method

3.1 Overview

Directly converting information-sparse PathCap and PathGen (Sun et al., 2024; 2025c) captions into VQA pairs degenerates into caption paraphrasing and conflates distinct diagnostic skills within a single training pool. We address both gaps with a three-stage pipeline over image–caption pairs $\mathcal{D} = \{(V_i, c_i)\}$, governed by the principle: *knowledge first, difficulty second, then track-conditioned rationales* (Fig. 2).

Stage I (Sec. 3.2) upgrades diagnostically eligible image–caption pairs into Verified Diagnostic Knowledge (VDK): a text-only disease gate filters eligible captions, retained samples are matched against a shared knowledge store $\mathcal{K}$, and unmatched cases undergo VLM-driven generation with text–visual auditing before updating $\mathcal{K}$.

Stage II (Sec. 3.3) proposes candidate question–answer pairs and partitions them into two difficulty levels, Easy and Hard, using an independent probe VLM. Easy items are solvable from the query image alone. Hard items are not; they are constructed along two unordered axes that name the missing evidence rather than a further difficulty rank: a knowledge axis that additionally requires VDK, and an image axis that is rescued only by an auxiliary contrastive neighbor.

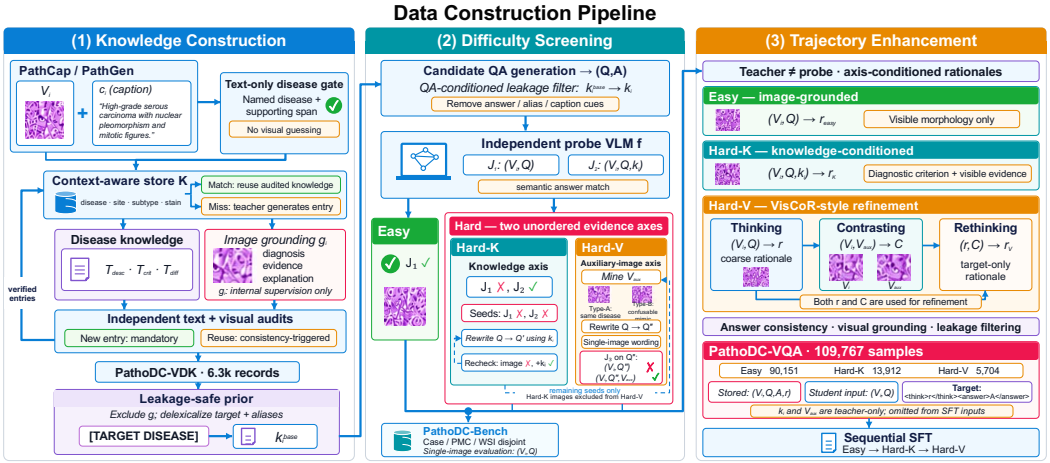


Figure 2: **Pathology VQA construction pipeline.** Stage I converts diagnostically eligible image–caption pairs into leakage-safe Verified Diagnostic Knowledge (VDK) and candidate VQA triples. Stage II uses probe tests $J_1/J_2/J_3$ to separate image-solvable Easy items from Hard items whose missing evidence is either VDK (Hard-K) or an auxiliary contrastive image (Hard-V); the two Hard axes are unordered. Stage III generates track-conditioned rationales—image-grounded for Easy, VDK-conditioned for Hard-K, and contrastively refined for Hard-V—and filters them before SFT. The resulting (V, Q, A, r) examples use only (V, Q) as model input; VDK and auxiliary images remain construction-time evidence.

Stage III (Sec. 3.4) augments each subset with axis-conditioned reasoning chains: image-grounded rationales for Easy, VDK-conditioned rationales for knowledge-axis Hard, and a VisCoR-style "thinking–contrasting–rethinking" distillation (Pan et al., 2026) for image-axis Hard. All rationales are filtered for answer consistency, visual grounding, and label/caption leakage before the subsets are integrated for supervised fine-tuning.

3.2 Verified Diagnostic Knowledge Construction

PathCap and PathGen cover many images, but their captions often record only stain, magnification, or experimental context, and neither corpus has a dedicated disease field. Diagnostic reasoning needs a different prior: what the disease is, how it differs from its mimics, and which morphological criteria apply to the current image. We therefore construct Verified Diagnostic Knowledge (VDK) from existing image–caption pairs (V_i, c_i), using audit filters in place of manual textbook curation. Field definitions and the consistency score are given in Appendix A; prompt templates are in Appendix D (Figs. 4–8).

Construction. VDK is built in five steps. (1) A text-only disease gate reads the caption and keeps a sample only when the caption explicitly names a disease and the supporting span is present in that caption. Generic findings such as "inflammation" or "lesion", method-centric captions without a disease entity, and technical identifiers mistaken for diagnoses are rejected. The image is withheld at this step so that visual guesses do not enter the store. (2) A VLM checks the retained disease name. Synonyms, abbreviations, and the canonical form must refer to the same disease; mismatches go to text audit and unresolved conflicts are discarded. The store key is not the disease string alone. It is the disease together with anatomical site, histological subtype or variant, and stain or imaging modality. Missing attributes may stay empty, but a known contextual conflict is never merged away. Each record stores audited aliases, a disease description, distinctions from one or two confusable mimics, and the key diagnostic criteria. (3) If that context has no record, the VLM writes the full disease-level entry and a separate image grounding used only for this case: the figure-level diagnosis, the visible morphology, and how that morphology supports the diagnosis. The disease-level entry is not committed until it passes verification. If a record already exists, the audited description and criteria are injected as fixed context, and the model adds only the case-specific differential and image grounding. Disease definitions therefore stay stable across samples, while image-specific

evidence is kept. (4) Every new disease-level entry is audited by an independent text LLM for entity consistency, professional wording, and compatibility with the caption. Image grounding is checked by an independent VLM: the claimed morphology must be visible in V_i, and the explanation must link that evidence to the diagnostic criteria. Reused records are not accepted on self-reported confidence. They are routed by multi-sample consistency (Wang et al., 2023): whether the diagnosis agrees across samples, and whether the morphological evidence and explanation agree with each other. Low consistency, entity conflicts, missing fields, or schema violations trigger visual audit; inconsistencies in the disease name, aliases, caption, or context key also trigger text audit. Failure of either required audit discards the record. Sampling count, mixture weight, and audit routing are specified in the Reproducibility Statement and Appendix F. (5) The figure-level diagnosis and its explanation are kept for question writing, answer checks, and rationale supervision. They are never placed in the knowledge-axis Hard input. The prior exposed downstream, denoted k_i, replaces the disease name and its aliases by `[TARGET DISEASE]`. After Stage II proposes (Q, A), a second filter removes spans that copy the answer, an alias, the figure-level diagnosis, or a caption cue. A sample is discarded if the answer can still be copied from the remaining text, or if sanitization removes the diagnostic content. Throughout the paper, k_i is this leakage-safe prior and must be used together with the image.

Role of the design. The context key stops knowledge from being merged across site, subtype, or stain. Auditing before insertion makes later reuse follow a stable disease definition rather than one figure's wording. Delexicalization forces the model to align diagnostic criteria with visible morphology.

3.3 Difficulty-Controlled VQA

The VDK constructed in Sec. 3.2 provides a diagnostic prior, but it does not say whether a training item is decidable from the query image alone. Items that fail this test are harder than image-solvable ones, yet the extra evidence they need is not itself a difficulty order: a leakage-safe knowledge prior and an auxiliary contrastive image are different kinds of missing evidence, and neither is uniformly harder than the other. We therefore use two difficulty levels, Easy and Hard, and construct Hard along two unordered axes. Admission is defined by an independent probe VLM f. The correctness function $s(f(\cdot), A)$ equals 1 when an answer judge accepts the probe output as a semantic match to the reference answer A, and equals 0 otherwise. It is not exact string equality, nor a canonical-name check by itself. The probe and semantic answer judge used in J_3 are specified in the Reproducibility Statement; the judge template is given in Appendix F (`qwen-max`). Three tests are used below and summarized in Appendix G (Table 5): $J_1 = s(f(V, Q), A)$ on the query image alone, $J_2 = s(f(V, Q, k), A)$ after k is added, and J_3 compares the single-image and auxiliary-image probes on the rewritten query.

Starting from the VDK pool, a VLM proposes candidate (Q, A) pairs conditioned on (V_i, c_i, k_i). After text-level coherence checks and deduplication, Alg. 1 (Appendix C) assigns each candidate by the probe f. Phase 1 admits an item to Easy $\mathcal{E}$ when $s(f(V_i, Q), A) = 1$. If the image-only probe fails and $s(f(V_i, Q, k_i), A) = 1$, the same phase admits it to knowledge-axis Hard $\mathcal{H}_{\mathrm{K}}$. Only items that fail both probes enter the seed pool $\mathcal{S}$. Phase 2 rewrites seeds whose image is not yet represented in $\mathcal{H}_{\mathrm{K}}$; a successful rewrite is added to $\mathcal{H}_{\mathrm{K}}$ and removed from $\mathcal{S}$. Phase 3 builds image-axis Hard $\mathcal{H}_{\mathrm{V}}$ only from seeds that remain in $\mathcal{S}$, so any sample already in $\mathcal{H}_{\mathrm{K}}$ is excluded. On f, both Hard axes fail image-only answering and succeed after their own extra cue is added.

Easy: single-image solvable VQA. Easy targets morphological findings that the probe decides from the query image alone. An item is admitted to $\mathcal{E}$ when $J_1 = 1$. The formal test is Phase 1 of Alg. 1. These samples supervise foundational visual perception. They do not require k_i or an auxiliary image at inference, although VDK still guides question authorship during construction. Keeping them separate from Hard prevents image-solvable items from diluting supervision that depends on extra evidence.

Hard, two construction axes. On the probe f, every Hard item has $J_1 = 0$ and then succeeds after one extra cue is added. The cue is either the leakage-safe prior k_i or an auxiliary image V_{aux}.

Knowledge-axis Hard. This axis targets criteria and differential questions for which f fails on (V_i, Q) and succeeds once k_i from Sec. 3.2 is added, that is $J_1 = 0$ and $J_2 = 1$. The test records the probe outcome. It does not claim that every such question is unanswerable without k_i. Phase 1 already admits candidates that meet this test. Phase 2 covers images that fail both (V_i, Q) and (V_i, Q, k_i) on f: it rewrites Q into Q' toward differential criteria in k_i, with at most R_{K} retries. The rewrite is kept only when f still fails on (V_i, Q') and succeeds on (V_i, Q', k_i). Accepted images are removed from $\mathcal{S}$ and do not enter Phase 3. This axis trains knowledge-conditioned reasoning and does not use an auxiliary image.

Image-axis Hard. The complementary axis supplies a visual reference instead of text knowledge. Phase 3 uses only seeds still in $\mathcal{S}$, which excludes every sample already admitted to $\mathcal{H}_{\mathrm{K}}$. For each remaining seed, we mine up to six nearest neighbors with pathology embeddings (Sun et al., 2024; Zhang et al., 2026) and keep two contrastive types:

- **Type-A (intra-class):** the same disease under a different morphological presentation, so the neighbor confirms visual variation.
- **Type-B (inter-class):** a confusable look-alike disease, so the neighbor forces a fine-grained distinction.

For each pair (V_i, V_{aux}), we rewrite Q as a single-image query. The wording never refers to a second image, so V_{aux} is construction-time evidence rather than a deployment input. Type-B rewrites may include a trap that leans toward the mimic, while the gold answer remains the query disease. An item enters $\mathcal{H}_{\mathrm{V}}$ only when, on f, the rewritten query fails from the image alone and succeeds after V_{aux} is added. That pair of outcomes is J_3, measured on Q'' rather than on the Phase 1 question. This is a probe outcome, not a claim that the question cannot be answered without the auxiliary image. At inference the question remains single-image: Stage III distills the contrast into its rationale, so deployment does not receive V_{aux}.

The released PathoDC-VQA pool contains 90,151 Easy, 13,912 knowledge-axis Hard, and 5,704 image-axis Hard questions (N=109,767; Appendix G, Table 5). Easy is disjoint from both Hard axes. Held-out question-quality evidence is reported in Appendix G (Table 6).

3.4 Reasoning-Oriented SFT with Axis-Conditioned Rationales

Sec. 3.3 separates items by which extra evidence the probe needs: none for Easy, k_i for knowledge-axis Hard, and V_{aux} for image-axis Hard. Answer-only supervision drops that distinction. We therefore attach a chain-of-thought rationale r (Wei et al., 2022) whose generation protocol follows the evidence axis, so the student is trained on how A is justified.

Shared generation protocol. A teacher VLM, distinct from the probe f, drafts each rationale. For every axis we sample two trajectories: reason-then-answer, and justify-the-fixed-answer. An answer judge keeps a candidate only when its conclusion matches A. The same filter removes caption copying, VDK answer leakage, and meta-phrases such as "the text says" or "according to the caption". Drafts that fail either check are not added to the training pool. The teacher may read axis-specific context while writing r, as defined below. That context is not stored in the training set. For Easy, Hard-K, and Hard-V, each SFT example is the tuple (V_i, Q, A, r). At evaluation the model is given only (V_i, Q).

Easy: single-image rationales. The teacher drafts r_{easy} from (V_i, Q), citing only morphological findings visible in V_i. The rationale does not mention k_i or an auxiliary image. The SFT example is $(V_i, Q, A, r_{\mathrm{easy}})$.

Knowledge-axis Hard: VDK-conditioned rationales. The teacher writes r_{K} from (V_i, Q, k_i). The rationale must cite a diagnostic criterion or a look-alike distinction in the leakage-safe prior and bind that statement to a finding visible in V_i, so it does not become a paraphrase of k_i. We keep the draft only when it is consistent with A and the probe still records $J_1 = 0$ on (V_i, Q). That check is the probe outcome from Sec. 3.3; it does not claim that the question is unanswerable without k_i. k_i is available only to the teacher. The SFT example is $(V_i, Q, A, r_{\mathrm{K}})$ and does not contain k_i. The

criterion has to be written into r_{K}, because the training tuple and the PathoDC-Bench prompt both omit k_i (Fig. 2).

Image-axis Hard: contrastive refinement. Each $\mathcal{H}_{\mathrm{V}}$ item was admitted with an auxiliary neighbor V_{aux}, while its question wording stays single-image. The teacher follows the "thinking–contrasting–rethinking" protocol: it drafts from V_i alone (*thinking*), compares that draft with V_{aux} under the neighbor type (*contrasting*), and rewrites the comparison into a target-only rationale r_{V} (*rethinking*). V_{aux} is available only to the teacher. The SFT example is $(V_i, Q, A, r_{\mathrm{V}})$ and does not contain V_{aux}. The training mixture and likelihood are in Appendix E.

4 Experiments

We test PathoDC on public pathology VQA (Quilt-VQA (Seyfioglu et al., 2024) and PathVQA (He et al., 2021)) and a taxonomy-aligned case-disjoint holdout (PathoDC-Bench), followed by stage-wise analysis. These settings ask whether the curriculum transfers, where its largest gains occur, and whether Hard-K and Hard-V produce the intended evidence-specific skills. Additional atlas-style transfer results on PathMMU (Sun et al., 2025b) are reported in Appendix I.

4.1 Experimental Setup

Models and training. **PathoDC-3B** and **PathoDC-7B** start from the official Qwen2.5-VL Instruct 3B/7B checkpoints (Bai et al., 2025) and undergo sequential Easy→Hard-K→Hard-V SFT on PathoDC-VQA (N=109,767; 90,151 Easy, 13,912 Hard-K, and 5,704 Hard-V). Training uses no reinforcement learning. The two scales share data order, optimization, and image preprocessing; full settings are in Appendix F. We also evaluate checkpoints after each curriculum stage.

Comparison models. We compare against general-purpose VLMs: DeepSeek-VL2-Tiny (Wu et al., 2024), Qwen2.5-VL-3B/7B (Bai et al., 2025), InternVL2-8B (Chen et al., 2024b), InternVL2.5-8B (Chen et al., 2025), InternVL3-8B (Zhu et al., 2025), LLaVA-OneVision-7B (Li et al., 2024), and Llama-3.2-Vision-11B-cot (Xu et al., 2025); and pathology VLMs: PathGen-LLaVA-13B (Sun et al., 2025c), Quilt-LLaVA-v1.5-7B (Seyfioglu et al., 2024), HuatuoGPT-Vision-7B (Chen et al., 2024a), and Patho-R1-3B/7B (Zhang et al., 2026). Official checkpoints are evaluated as released. Patho-R1 uses 550k SFT examples plus RL and is therefore a data-scale reference, not a same-recipe ablation.

Benchmarks and metrics. Quilt-VQA and PathVQA report Closed, Open, and Overall accuracy using the VLMEvalKit ShortQA protocol with `qwen3.8-max` as judge. PathoDC-Bench contains 1,000/500/300 case-disjoint Easy/Hard-K/Hard-V items with no training-case, PMC, or WSI overlap. Besides split accuracy, we report sample-weighted (WAvg), macro (Avg), and 1:2:2 difficulty-weighted (DWS) averages. Judge details and quality audits are in Appendices F and G; the PathMMU protocol is in Appendix I.

4.2 Public Pathology VQA

We use identical Closed/Open/Overall metrics and judge settings on Quilt-VQA and PathVQA; Table 1 compares 3B and 7B/8B models.

Results. PathoDC-3B improves its Instruct initialization by 0.39/1.67 points on Quilt-VQA/PathVQA Overall, leading the compared 3B models on PathVQA. PathoDC-7B gains 14.91/1.93 points and leads all models on Quilt-VQA Overall (36.50), while HuatuoGPT-Vision remains stronger on PathVQA (35.81 vs. 34.50) and Patho-R1-3B remains stronger on Quilt-VQA. The evidence supports transfer, not universal SOTA.

4.3 Taxonomy-Aligned Diagnostic Evaluation

PathoDC-Bench directly tests the three construction tracks against the Instruct backbones and public pathology VLMs (Table 2).

Table 1: Pathology VQA accuracy (%). PathoDC-3B / PathoDC-7B: sequential Easy→Hard-K→Hard-V finals on **PathoDC-VQA** (N=109,767). Best / second-best: **bold** / underlined.

	Model	Quilt-VQA			PathVQA		
		Closed↑	Open↑	Overall↑	Closed↑	Open↑	Overall↑
General VLM	Deepseek-VL2-Tiny	42.32	6.31	15.94	44.28	4.30	24.35
	Qwen2.5-VL-3B	**62.68**	10.96	24.79	61.46	5.04	33.34
Pathology VLM	Patho-R1-3B	60.64	**24.68**	**34.29**	53.91	**9.61**	31.83
	PathoDC-3B	60.35	12.34	25.18	**61.87**	7.98	**35.01**
General VLM	InternVL2-8B	52.77	12.13	22.99	55.20	7.48	31.42
	InternVL2.5-8B	45.48	13.40	21.98	38.16	8.52	23.38
	InternVL3-8B	48.10	14.04	23.15	44.50	11.13	27.87
	LLaVA-Onevision-7B	21.67	5.30	9.68	52.16	0.68	26.50
	Llama-3.2V-11B-cot	51.85	14.10	24.19	56.23	3.73	30.06
	Qwen2.5-VL-7B	40.23	14.79	21.59	57.48	7.51	32.57
Pathology VLM	PathGen-LLaVA-13B	51.90	12.23	22.84	57.53	3.83	30.76
	Quilt-LLaVA-v1.5-7B	31.49	9.57	15.43	33.32	2.20	17.81
	HuatuoGPT-Vision-7B	43.73	12.34	20.73	**62.22**	9.23	**35.81**
	Patho-R1-7B	61.81	25.64	35.31	57.89	**11.28**	34.65
	PathoDC-7B	**63.50**	**26.70**	**36.50**	62.10	6.80	34.50

Results. Both PathoDC scales lead their model groups on every split. WAvg rises by 36.39/42.34 points over the 3B/7B Instruct backbones and by 9.89/21.17 points over Patho-R1, despite using 109,767 SFT examples without RL. The largest gains therefore occur on the taxonomy-aligned holdout, not uniformly on public VQA or MCQ.

Table 2: PathoDC-Bench accuracy (%) by evidence track. WAvg follows split size, Avg is the macro average, and DWS weights Easy/Hard-K/Hard-V as 1:2:2. Best / second-best within each model-scale group: **bold** / underlined; the 13B model is listed separately.

	Model	PathoDC-Bench					
		Easy↑	Hard-K↑	Hard-V↑	WAvg↑	Avg↑	DWS↑
General VLM	Deepseek-VL2-Tiny	–	–	–	–	–	–
	Qwen2.5-VL-3B	22.70	67.20	24.00	35.28	37.97	41.02
Pathology VLM	Patho-R1-3B	54.30	80.80	55.00	61.78	63.37	65.18
	PathoDC-3B	**66.70**	**83.40**	**68.67**	**71.67**	**72.92**	**74.17**
General VLM	InternVL2-8B	52.20	70.40	41.00	55.39	54.53	55.00
	InternVL2.5-8B	47.20	45.40	42.67	45.94	45.09	44.67
	InternVL3-8B	48.00	66.40	41.67	52.06	52.02	52.83
	LLaVA-Onevision-7B	47.96	46.32	53.21	48.38	49.16	49.40
	Llama-3.2V-11B-cot	50.35	45.23	47.95	48.53	47.84	47.34
	Qwen2.5-VL-7B	40.50	26.80	40.67	36.72	35.99	35.09
Pathology VLM (13B)	PathGen-LLaVA-13B	42.50	33.80	31.67	38.28	35.99	34.69
Pathology VLM	Quilt-LLaVA-v1.5-7B	25.40	44.20	27.33	30.94	32.31	33.69
	HuatuoGPT-Vision-7B	43.20	73.00	42.67	51.39	52.96	54.91
	Patho-R1-7B	51.10	77.20	48.33	57.89	58.88	60.43
	PathoDC-7B	**75.70**	**87.80**	**75.67**	**79.06**	**79.72**	**80.53**

4.4 Curriculum-Stage Analysis

Table 3 compares caption pretraining, Easy, Easy→Hard-K, and the full curriculum; stage curves and additional comparisons are in Appendix J.

Results. Easy SFT supplies most of the broad gain: WAvg rises from 35.28 to 73.06 at 3B and from 36.72 to 78.56 at 7B. Easy→Hard-K then peaks on Hard-K (88.40/90.20), while the full curriculum peaks on Hard-V (68.67/75.67), supporting evidence-axis specialization. The final model loses 4.80/1.50 Easy points and 5.00/2.40 Hard-K points, so it trails Easy-only in 3B WAvg but attains the best 7B WAvg (79.06). Caption-pretraining-only is below Instruct on public VQA and PathoDC-Bench; recovery after diagnostic VQA shows that extra caption exposure alone does not explain the gain. The remaining forgetting motivates replay or joint-mixture training.

Table 3: Stage-wise accuracy (%) on public VQA, PathMMU, and PathoDC-Bench. Intermediate rows stop after caption pretraining (CPT), Easy, or Easy→Hard-K. Best / second-best within 3B or 7B: **bold** / underlined.

Variants	Quilt-VQA↑			PathVQA↑			PathMMU All↑			PathoDC-Bench↑				
	Closed	Open	All	Closed	Open	All	tiny	test	val	Easy	Hard-K	Hard-V	WAvg	Avg
3B														
PathoDC-Easy-3B	62.10	13.30	26.34	**62.75**	6.44	34.68	51.14	48.54	**48.47**	**71.50**	83.60	60.67	**73.06**	71.92
PathoDC-Hard-K-3B	**67.06**	**13.40**	**27.75**	57.59	6.50	32.13	52.12	49.11	47.93	67.70	**88.40**	64.00	72.83	**73.37**
PathoDC-3B	60.35	12.34	25.18	61.87	**7.98**	**35.01**	**62.27**	**57.43**	43.60	66.70	83.40	**68.67**	71.67	72.92
7B														
PathoDC-CPT-7B	25.07	10.32	14.26	43.56	3.83	23.75	55.16	48.93	49.91	18.40	18.40	12.00	17.33	16.27
PathoDC-Easy-7B	62.97	25.85	35.78	61.81	6.35	34.17	55.92	53.93	54.41	**77.20**	88.40	66.67	78.56	77.42
PathoDC-Hard-K-7B	60.64	21.91	32.27	57.86	5.73	31.87	54.51	51.89	54.23	71.50	**90.20**	66.00	75.78	75.90
PathoDC-7B	**63.50**	**26.70**	**36.50**	**62.10**	**6.80**	**34.50**	**68.71**	**62.90**	**54.95**	75.70	87.80	**75.67**	**79.06**	**79.72**

4.5 Component and Ordering Ablations

Table 4 reports measured controls for VDK, rationale supervision, difficulty assignment, the training mixture, auxiliary-image selection, and curriculum order. Every entry is an experimental result; no row contains projected or expected values.

Table 4: Measured ablations of PathoDC supervision. Non-final controls are evaluated at 7B; matched controls keep data size and optimizer-update budget fixed. Shuffled mixture uses Quilt caption pretraining plus one shuffled pass and is reported only as a recipe-level comparison. WAvg follows the PathoDC-Bench split sizes.

Variant	Quilt All↑	PathVQA All↑	Easy↑	Hard-K↑	Hard-V↑	WAvg↑
PathoDC-3B	25.18	35.01	66.70	83.40	68.67	71.67
PathoDC-7B	36.50	34.50	75.70	87.80	75.67	79.06
−VDK	32.40	33.10	70.40	72.10	64.80	69.94
Answer-only	34.20	33.80	73.50	85.20	62.40	74.90
Generic rationale	35.10	34.00	74.20	86.10	68.50	76.56
Random difficulty split	33.80	33.40	68.20	78.50	64.00	70.36
Shuffled mixture	35.23	33.16	74.72	86.80	73.76	77.91
Random auxiliary image	35.40	34.20	75.10	87.20	58.30	75.66
Reverse order	34.60	34.10	76.80	80.20	63.50	75.53

Results. Removing VDK yields the largest Hard-K drop (-15.70), while answer-only and generic rationales reduce Hard-V by 13.27/7.17 points. Under matched data and budgets, randomizing difficulty lowers WAvg by 8.70 points, reversing the order reduces Hard-K/Hard-V by 7.60/12.17, and random auxiliary images reduce Hard-V by 17.37. The separately specified shuffled recipe is also lower (Easy/Hard-K/Hard-V: $-0.98/-1.00/-1.91$; WAvg: -1.15; public Overall: $-1.27/-1.34$), but it changes caption pretraining and data exposure, so the gap is not attributed to ordering alone.

5 Conclusion

We address the gap between information-sparse pathology image–caption corpora and supervision for diagnostic reasoning by constructing layered training data: Verified Diagnostic Knowledge first, then an Easy/Hard split whose Hard items are built from knowledge and auxiliary-image axes, and finally axis-conditioned rationales for SFT. The resulting PathoDC models show that Easy SFT drives most transfer, the two Hard axes are skill-specific, and compact **109,767** SFT can exceed a **550k** SFT+RL baseline on PathoDC-Bench, the holdout that matches the curriculum. Public open-ended VQA and atlas-style PathMMU remain mixed, so the claim is taxonomy-aligned diagnostic supervision—not universal SOTA from less data.

A current limitation is that construction and scoring still rely on teacher VLMs and LLM-as-judge protocols, and the independent Hard-K probe shows many items remain solvable without k. The human-audit protocol is in Appendix H. Extending the curriculum to whole-slide contexts and closing the loop with reinforcement learning on the same track-aligned format are natural next steps.

ETHICS AND PRIVACY STATEMENT

This work develops a curriculum for training pathology vision–language models from publicly released image–caption corpora (PathCap and PathGen) and evaluates on public benchmarks; we do not collect new patient-identifiable clinical records, and use of these resources should follow their original licenses and institutional data-use policies. Because PathoDC improves automated diagnostic-style answering, a primary societal risk is over-reliance or misuse as an unsupervised clinical decision system; the method is intended for research and educational model development under qualified pathologist oversight, not for standalone diagnosis or triage. Synthetic knowledge and LLM-as-judge scoring can also propagate dataset bias (e.g., disease, stain, or site imbalance) or plausible but incorrect rationales, so fairness, safety, and accountability require human review before any downstream deployment. Overall, the intended benefit is more transparent, difficulty-controlled supervision for pathology multimodal research, balanced against the above misuse and error-propagation risks.

AI USE STATEMENT

Generative AI tools were used solely for language polishing and improving the clarity and fluency of the manuscript. All suggested revisions were reviewed and approved by the authors, who take full responsibility for the final content.

REPRODUCIBILITY STATEMENT

Construction procedures, admission rules $J_1/J_2/J_3$, prompts, and the SFT schedule are in Sec. 3.2–3.4 and Appendices A–D. The values below are the ones used to build PathoDC-VQA and to train PathoDC-3B / PathoDC-7B. We will release the construction scripts, LLaMA-Factory configs, and evaluation commands with the camera-ready version.

Models (DashScope OpenAI-compatible API). Disease gate: `qwen-plus`. Teacher VLM (VDK generation, question writing, Easy/Hard-K rationales): `qwen-vl-max`. Probe VLM f ($J_1/J_2/J_3$ image answering): `qwen-vl-plus`. Text auditor and semantic answer-match judge $s(\cdot, A)$: `qwen-max`. Independent quality probe: `qwen-vl-plus` (seed 42, n=40 per track). Public Quilt-VQA / PathVQA judge: `qwen3.8-max` with the VLMEvalKit ShortQA template (Appendix F), temperature 0, `max_tokens`=2048. PathMMU letter extraction fallback: `qwen-vl-max`.

Stage I constants. The released reuse path injects audited $T_{\text{desc}}/T_{\text{crit}}$ on a cache hit and samples case-specific fields M=5 times. Eq. 8 uses threshold τ=0.7: a reused record is sent to visual audit when C_i<0.7, or when it has an entity conflict, a missing field, or a schema violation. Length bounds are 80–2500 characters; a structurally invalid record is regenerated once and discarded if it still fails. Neighbor mining for Hard-V uses Patho-CLIP-L (`ViT-L-14`, `WenchuanZhang/Patho-CLIP-L`) cosine similarity, with $n_{\text{same-similar}}$=1, $n_{\text{same-dissimilar}}$=2, $n_{\text{confusable}}$=3 (six neighbors). Construction seed is 42.

Stage II retry budgets and decoding. R_{K}=R_{V}=2 (at most 1+2=3 rewrites). Temperature: gate/audit 0.0; VDK generation 0.2; Easy questions 0.1; Hard-K rewrites 0.35; Hard-V first attempt 0.45 and retries 0.75. `top_p` is left at the API default. Hard-V JSON calls use `max_tokens`=1400. Holdout split ratio 0.1, seed 42.

SFT (PathoDC-3B / PathoDC-7B). Full-parameter SFT of the language model; vision tower and multimodal projector are frozen (not LoRA). Optimizer: `adamw_torch_fused`, β=(0.9, 0.999), ε=10^{-8}. Learning rate 1×10^{-5}, cosine decay, warmup ratio 0.03, `bf16`, DeepSpeed ZeRO-3, cutoff 4096 tokens. Images use Qwen2.5-VL dynamic resolution with `image_max_pixels`=262144 ($\approx$ 512×512). Each curriculum stage runs 3 epochs, seed 42. PathoDC-3B: 7×A100-80GB, per-device batch 1, gradient accumulation 8, global batch 56; steps 4830 / 537 / 186 (Easy / Hard-K / Hard-V); wall-clock 24.9 / 7.4 / 2.0 hours ($\approx$ 34 hours total). PathoDC-7B without caption pretrain:

7×A100-80GB, per-device batch 2, accumulation 8, global batch 112; steps 2415 / 270 / 93; wall-clock 24.8 / 5.2 / 3.5 hours. The 7B Quilt-pretrain ablation uses 8×A100-80GB, per-device batch 4, global batch 256, and a 3-epoch caption warmup on 300k Quilt pairs (3516 steps, 23.0 hours) before the same Easy→Hard-K→Hard-V schedule.

Ablation controls. The −VDK, Generic rationale, Random difficulty split, and Reverse order rows in Table 4 are equal-data, equal-budget controls at 7B. They use the same number of training examples, total optimizer updates, optimizer settings, image preprocessing, and evaluation protocol as the full PathoDC-7B run. The random split preserves the original 90,151/13,912/5,704 stage cardinalities, and the reverse-order run retains the same three stage pools and three epochs per stage. The Shuffled mixture row is not included in this matched set: it uses the Quilt-pretrained initialization and one pass over the combined pool, so it is reported only as a recipe-level comparison.

Pointers. Question-quality probes and the human-audit protocol: Appendix G. PathMMU subset scores: Appendix I. Stage-variant results: Table 3 and Appendix J.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 41–48. ACM, 2009. doi: 10.1145/1553374.1553380. URL https://doi.org/10.1145/1553374.1553380.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. HuatuoGPT-Vision, towards injecting medical visual knowledge into multimodal LLMs at scale, 2024a. URL https://arxiv.org/abs/2406.19280.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to GPT-4V? closing the gap to commercial multimodal models with open-source suites, 2024b. URL https://arxiv.org/abs/2404.16821.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Xuehai He, Zhuo Cai, Wenlan Wei, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Towards visual question answering on pathology images. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 708–718. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-short.90. URL https://aclanthology.org/2021.acl-short.90/.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer, 2024. URL https://arxiv.org/abs/2408.03326.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. In *Advances in Neural Information Processing Systems*, volume 36, pp. 28541–28564, 2023. doi: 10.52202/075280-1240. URL https://doi.org/10.52202/075280-1240.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Ivy Liang, Tong Ding, Guillaume Jaume, Igor Odintsov, Long Phi Le, Georg Gerber, Anil V. Parwani, Andrew Zhang, and Faisal Mahmood. A visual-language foundation model for computational pathology. *Nature Medicine*, 30(3):863–874, 2024a. doi: 10.1038/s41591-024-02856-4. URL https://doi.org/10.1038/s41591-024-02856-4.

Ming Y. Lu, Bowen Chen, Drew F. K. Williamson, Richard J. Chen, Melissa Zhao, Aaron K. Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, Amr Soliman, Chengkuan Chen, Tong Ding, Judy J. Wang, Georg Gerber, Ivy Liang, Long Phi Le, Anil V. Parwani, Luca L. Weishaupt, and Faisal Mahmood. A multimodal generative AI copilot for human pathology. *Nature*, 634 (8033):466–473, 2024b. doi: 10.1038/s41586-024-07618-3. URL https://doi.org/10.1038/s41586-024-07618-3.

Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Cyril Zakka, Yash Dalmia, Eduardo Pontes Reis, Pranav Rajpurkar, and Jure Leskovec. Med-Flamingo: a multimodal medical few-shot learner, 2023. URL https://arxiv.org/abs/2307.15189.

Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning, 2025. URL https://arxiv.org/abs/2502.19634.

Zhiyu Pan, Yizheng Wu, Jiashen Hua, Junyi Feng, Shaotian Yan, Bing Deng, Zhiguo Cao, and Jieping Ye. Through the lens of contrast: Self-improving visual reasoning in VLMs. In *The Fourteenth International Conference on Learning Representations (ICLR)*, 2026. doi: 10.48550/arXiv.2603.02556. URL https://arxiv.org/abs/2603.02556.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, et al. MedGemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Mehmet Saygin Seyfioglu, Wisdom O. Ikezogwo, Fatemeh Ghezloo, Ranjay Krishna, and Linda Shapiro. Quilt-LLaVA: Visual instruction tuning by extracting localized narratives from open-source histopathology videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13183–13192, 2024. doi: 10.1109/CVPR52733.2024.01252. URL https://doi.org/10.1109/CVPR52733.2024.01252.

Yuxuan Sun, Chenglu Zhu, Sunyi Zheng, Kai Zhang, Lin Sun, Zhongyi Shui, Yunlong Zhang, Honglin Li, and Lin Yang. PathAsst: A generative foundation AI assistant towards artificial general intelligence of pathology. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5034–5042, 2024. doi: 10.1609/aaai.v38i5.28308. URL https://ojs.aaai.org/index.php/AAAI/article/view/28308.

Yuxuan Sun, Yixuan Si, Chenglu Zhu, Xuan Gong, Kai Zhang, Pingyi Chen, Ye Zhang, Zhongyi Shui, Tao Lin, and Lin Yang. CPath-Omni: A unified multimodal foundation model for patch and whole slide image analysis in computational pathology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10360–10371, 2025a. doi: 10.1109/CVPR52734.2025.00969. URL https://doi.org/10.1109/CVPR52734.2025.00969.

Yuxuan Sun, Hao Wu, Chenglu Zhu, Sunyi Zheng, Qizi Chen, Kai Zhang, Yunlong Zhang, Dan Wan, Xiaoxiao Lan, Mengyue Zheng, Jingxiong Li, Xinheng Lyu, Tao Lin, and Lin Yang. PathMMU: A massive multimodal expert-level benchmark for understanding and reasoning in pathology. In *Computer Vision – ECCV 2024*, volume 15120 of *Lecture Notes in Computer Science*, pp. 56–73. Springer, 2025b. doi: 10.1007/978-3-031-73033-7_4.

Yuxuan Sun, Yunlong Zhang, Yixuan Si, Chenglu Zhu, Zhongyi Shui, Kai Zhang, Jingxiong Li, Xingheng Lyu, Tao Lin, and Lin Yang. PathGen-1.6M: 1.6 million pathology image-text pairs generation through multi-agent collaboration. In *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025c. URL https://openreview.net/forum?id=rFpZnn11gj.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Jianyu Wu, Hao Yang, Xinhua Zeng, Guibing He, Zhiyu Chen, Zihui Li, Xiaochuan Zhang, Yangyang Ma, Run Fang, and Yang Liu. PathVLM-R1: A reinforcement learning-driven reasoning model for pathology visual-language tasks, 2025. URL https://arxiv.org/abs/2504.09258.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. DeepSeek-VL2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL https://arxiv.org/abs/2412.10302.

Hanguang Xiao, Feizhong Zhou, Xingyue Liu, Tianqi Liu, Zhipeng Li, Xin Liu, and Xiaoxuan Huang. A comprehensive survey of large language models and multimodal large language models in medicine. *Information Fusion*, 117:102888, 2025. doi: 10.1016/j.inffus.2024.102888. URL https://doi.org/10.1016/j.inffus.2024.102888.

Guowei Xu, Peng Jin, Ziang Wu, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. LLaVA-CoT: Let vision language models reason step-by-step. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2087–2098, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Xu_LLaVA-CoT_Let_Vision_Language_Models_Reason_Step-by-Step_ICCV_2025_paper.html.

Fang Yan, Jianfeng Wu, Jiawen Li, Wei Wang, Jiaxuan Lu, Wen Chen, Zizhao Gao, Jianan Li, Hong Yan, Jiabo Ma, Minda Chen, Yang Lu, Qing Chen, Yizhi Wang, Xitong Ling, Xuenian Wang, Zihan Wang, Qiang Huang, Shengyi Hua, Mianxin Liu, Lei Ma, Tian Shen, Xiaofan Zhang, Yonghong He, Hao Chen, Shaoting Zhang, and Zhe Wang. PathOrchestra: A comprehensive foundation model for computational pathology with over 100 diverse clinical-grade tasks, 2025. URL https://arxiv.org/abs/2503.24345.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.48550/arXiv.2203.14465. URL https://arxiv.org/abs/2203.14465.

Kai Zhang, Rong Zhou, Eashan Adhikarla, Zhiling Yan, Yixin Liu, Jun Yu, Zhengliang Liu, Xun Chen, Brian D. Davison, Hui Ren, Jing Huang, Chen Chen, Yuyin Zhou, Sunyang Fu, Wei Liu, Tianming Liu, Xiang Li, Yong Chen, Lifang He, James Zou, Quanzheng Li, Hongfang Liu, and Lichao Sun. A generalist vision–language foundation model for diverse biomedical tasks. *Nature Medicine*, 30(11):3129–3141, 2024. doi: 10.1038/s41591-024-03185-2. URL https://doi.org/10.1038/s41591-024-03185-2.

Wenchuan Zhang, Penghao Zhang, Jingru Guo, Tao Cheng, Jie Chen, Shuwan Zhang, Zhang Zhang, Yuhao Yi, and Hong Bu. Patho-R1: A multimodal reinforcement learning-based pathology expert reasoner. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 28418–28426, 2026. doi: 10.1609/aaai.v40i33.40071. URL https://ojs.aaai.org/index.php/AAAI/article/view/40071.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A VDK Field Definitions and Consistency Score

This appendix records the symbols removed from Sec. 3.2. The construction procedure in the main text is unchanged.

Disease gate. A text-only gate reads the caption without the image,

$$\mathcal{G}(c_i) \mapsto (e_i, d_i^g, q_i^g, b_i^g), \tag{1}$$

where $e_i \in \{0, 1\}$ indicates an explicit disease, d_i^g is the surface form, q_i^g is the supporting caption span, and b_i^g is a short justification. A sample is kept only when $e_i = 1$, $d_i^g \neq \emptyset$, and q_i^g occurs in c_i.

Canonical name and context key. A VLM reads (V_i, c_i, d_i^g) and returns a confirmed name d_i^v. We require

$$\mathrm{Canon}(d_i^v) = \mathrm{Canon}(d_i^g), \tag{2}$$

where $\mathrm{Canon}(\cdot)$ maps surface forms and audited aliases to one disease. The store key is

$$\kappa_i = (d_i^{\mathrm{canon}}, a_i, s_i, m_i)\,, \tag{3}$$

with $d_i^{\mathrm{canon}} = \mathrm{Canon}(d_i^v)$, and a_i, s_i, m_i the site, subtype or variant, and stain or imaging modality. The shared store is

$$\mathcal{K} : \kappa \mapsto (\mathcal{A}_\kappa, \mathcal{T}_{desc}, \mathcal{T}_{diff}, \mathcal{T}_{crit})\,, \tag{4}$$

namely audited aliases, a professional description, distinctions from one or two mimics, and key diagnostic criteria.

Image grounding and internal record. Full generation also writes image-specific grounding

$$g_i = (\hat{d}_i, \mathcal{E}_i, \rho_i), \tag{5}$$

where $\hat{d}_i$ is the figure-level diagnosis, $\mathcal{E}_i$ is visual evidence, and ρ_i links that evidence to the diagnosis. The internal record

$$\mathcal{R}_i^{VDK} = (\kappa_i, \mathcal{A}_{\kappa_i}, \mathcal{T}_{desc}, \mathcal{T}_{diff}, \mathcal{T}_{crit}, g_i) \tag{6}$$

is not the input shown to the VQA model. In reuse mode, audited $\mathcal{T}_{desc}$ and $\mathcal{T}_{crit}$ stay fixed, and the model writes only the case-specific differential and g_i.

Reuse consistency. For a reused record, the generator is sampled M times,

$$g_i^{(m)} = \left(\hat{d}_i^{(m)}, \mathcal{E}_i^{(m)}, \rho_i^{(m)}\right), \qquad m = 1, \ldots, M. \tag{7}$$

Consistency is

$$C_i = \lambda C_i^{diag} + (1 - \lambda) C_i^{reason}, \tag{8}$$

$$C_i^{diag} = \frac{1}{M} \max_d \sum_{m=1}^{M} \mathbb{I}\Big[\mathrm{Canon}\Big(\hat{d}_i^{(m)}\Big) = d\Big]\,, \tag{9}$$

$$C_i^{reason} = \frac{2}{M(M-1)} \sum_{p<q} \mathrm{SemSim}\Big(\mathcal{E}_i^{(p)} \oplus \rho_i^{(p)},\, \mathcal{E}_i^{(q)} \oplus \rho_i^{(q)}\Big)\,, \tag{10}$$

where $\oplus$ denotes concatenation. C_i^{diag} is canonical-diagnosis agreement and C_i^{reason} is pairwise semantic agreement between evidence and explanation. The generator is sampled $M{=}5$ times. A reused record is sent to visual audit when $C_i{<}\tau$ with $\tau{=}0.7$, or when it has an entity conflict, a missing field, a schema violation, or inconsistent evidence. Text audit is added when the conflict concerns the disease entity, alias map, caption, or context key. Neighbor retrieval for Hard-V uses Patho-CLIP-L cosine similarity (Appendix F).

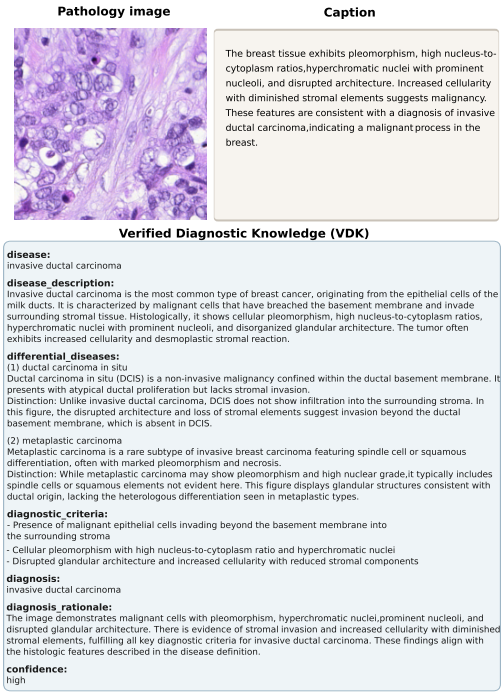


Figure 3: **Example of constructed VDK** (uid=`cd94ec4df6d04a05`): image, caption, and verified fields (`disease`, `diagnosis`, `confidence`, ...).

Leakage-safe prior. The delexicalized prior replaces the target disease and its audited aliases with `[TARGET DISEASE]`,

$$k_i^{base} = \mathrm{Delex}(\mathrm{Serialize}[\mathcal{T}_{desc}, \mathcal{T}_{diff}, \mathcal{T}_{crit}]\,; \mathcal{A}_{\kappa_i})\,. \tag{11}$$

After Stage II proposes (Q, A),

$$k_i^{Q,A} = \mathrm{Sanitize}\big(k_i^{base}; Q, A, c_i\big) \tag{12}$$

removes exact answers, aliases, equivalent answer statements, figure-level diagnoses, and caption cues. In the main text, k_i denotes $k_i^{Q,A}$.

B VDK Construction Example

Figure 3 shows one constructed VDK record: the source image and caption together with the verified diagnostic fields.

C Difficulty Construction Algorithm

Algorithm 1 is the complete three-phase procedure used in Sec. 3.3: image-only Easy admission, knowledge-axis Hard rewrites, and auxiliary-image Hard mining.

D VDK Generation and Verification Prompts

This appendix presents the prompt templates used for VDK construction. Bracketed variables are replaced with sample-specific inputs. All models are instructed to return valid JSON without additional commentary.

Algorithm 1 Two-level difficulty construction. Hard is realized on two unordered evidence axes.

Require: VDK pool $\mathcal{D} = \{(V_i, c_i, k_i)\}$; probe VLM f; correctness s; retry budgets $R_{\mathrm{K}}, R_{\mathrm{V}}$
Ensure: Easy $\mathcal{E}$; Hard axes $\mathcal{H}_{\mathrm{K}}$, $\mathcal{H}_{\mathrm{V}}$

Initialize $\mathcal{E}, \mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{V}}, \mathcal{S} \leftarrow \emptyset$
Phase 1: Easy vs. Hard by image sufficiency
for each $(V_i, c_i, k_i) \in \mathcal{D}$ **do**
  Propose candidate (Q, A) conditioned on (V_i, c_i, k_i)
  if $s(f(V_i, Q), A) = 1$ **then**
    $\mathcal{E} \leftarrow \mathcal{E} \cup \{(V_i, Q, A)\}$ ▷ Easy
  else if $s(f(V_i, Q, k_i), A) = 1$ **then**
    $\mathcal{H}_{\mathrm{K}} \leftarrow \mathcal{H}_{\mathrm{K}} \cup \{(V_i, Q, A, k_i)\}$ ▷ Hard, knowledge axis
  else
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{(V_i, Q, A)\}$ ▷ Hard seed
  end if
end for
Phase 2: Knowledge-axis Hard from seeds not yet in $\mathcal{H}_{\mathrm{K}}$
for each unique image V_i in $\mathcal{S}$ lacking representation in $\mathcal{H}_{\mathrm{K}}$ **do**
  for $r = 1$ **to** R_{K} **do**
    Rewrite $Q \to Q'$ targeting differential criteria in k_i
    if $s(f(V_i, Q'), A) = 0$ **and** $s(f(V_i, Q', k_i), A) = 1$ **then**
      $\mathcal{H}_{\mathrm{K}} \leftarrow \mathcal{H}_{\mathrm{K}} \cup \{(V_i, Q', A, k_i)\}$
      Remove all seeds of V_i from $\mathcal{S}$ ▷ admitted Hard-K skips Phase 3
      break
    end if
  end for
end for
Phase 3: Image-axis Hard; $\mathcal{H}_{\mathrm{K}}$ items are not in $\mathcal{S}$
for each remaining seed $(V_i, Q, A) \in \mathcal{S}$ **do**
  Mine V_{aux}: Type-A (same disease, different morphology) or Type-B (confusable look-alike)
  for each neighbor V_{aux} **and** $r = 1$ **to** R_{V} **do**
    Rewrite $Q \to Q''$ (strict single-image wording)
    if $s(f(V_i, Q''), A) = 0$ **and** $s(f(V_i, Q'', V_{\mathrm{aux}}), A) = 1$ **then**
      $\mathcal{H}_{\mathrm{V}} \leftarrow \mathcal{H}_{\mathrm{V}} \cup \{(V_i, V_{\mathrm{aux}}, Q'', A)\}$; **break**
    end if
  end for
end for
return $\mathcal{E}, \mathcal{H}_{\mathrm{K}}, \mathcal{H}_{\mathrm{V}}$

Deterministic post-filters. In addition to the model-based audits above, we apply deterministic checks for JSON validity, required-field completeness, UID duplication, length bounds, placeholder preservation, forbidden meta-references, and exact answer or alias overlap. A record that fails structural validation is regenerated once and discarded if the violation persists. Length bounds are 80–2500 characters. The remaining construction and SFT values are listed in the Reproducibility Statement and Appendix F.

E SFT Mixture and Objective

The instruction set $\mathcal{D}_{\mathrm{SFT}}$ is the full PathoDC-VQA pool: 90,151 Easy, 13,912 Hard-K, and 5,704 Hard-V (N=109,767; Table 5). Each training example is (V_i, Q, A, r), with neither k_i nor V_{aux} included. In the likelihood, $X = (V_i, Q)$ is the prompt and Y is the supervised continuation: the rationale inside `<think>...</think>`, then the short answer inside `<answer>...</answer>`. Evaluation prompts contain only (V_i, Q). We minimize the standard autoregressive negative log-likelihood:

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\mathbb{E}_{(X,Y)\sim\mathcal{D}_{\mathrm{SFT}}} \left[\sum_{t=1}^{|Y|} \log P_\theta(y_t \mid X, y_{<t}) \right]. \tag{13}$$

By generating the rationale before the short answer, this stage teaches evidence-consistent diagnostic reasoning on Easy and on both Hard axes, and leaves a structured target for later reinforcement learning.

Caption-Only Disease Gate

```
SYSTEM:
You are a pathology caption eligibility auditor. Determine eligibility
using ONLY the supplied caption. Do not infer a diagnosis from an image
or from external patient context.

A caption is eligible only when it explicitly names a specific disease
or diagnostically recognized disease entity.

Reject the sample if:

1. it contains only generic findings such as "inflammation", "lesion",
   "tumor", or "abnormal tissue" without a specific disease entity;
2. a diagnosis would require interpreting an unseen image;
3. it describes only staining, magnification, experimental procedures,
   dataset identifiers, or acquisition settings;
4. a TCGA identifier, stain name, gene name, or technical term is being
   misinterpreted as a diagnosis;
5. the disease name cannot be supported by an exact caption span.

Return:
{
  "is_explicit": true or false,
  "disease_surface": string or null,
  "canonical_candidate": string or null,
  "evidence_span": exact substring from the caption or null,
  "decision_reason": one brief sentence
}

USER:
Caption: [CAPTION]
```

Figure 4: Prompt template for the caption-only disease gate.

F Implementation Details

This appendix records the judge template and a compact copy of the values listed in the Reproducibility Statement.

Public-VQA judge. Quilt-VQA and PathVQA use VLMEvalKit ShortQA with `qwen3.8-max` (DashScope). The English template is:

> You are an AI assistant tasked with evaluating whether a model's response correctly answers a given visual-language question. The response is correct if it has the same meaning as the ground truth, even if phrased differently, or provides additional relevant details without altering that meaning. It is wrong if it contradicts the ground truth, misses essential information, or includes additional incorrect information. Return `Correctness: yes/no` and a brief `Reason`.

Decoding uses temperature 0 and `max_tokens`$=2048$. A failed parse is retried up to three times with temperature $0.5\times i$.

Construction and SFT (copy of the released settings). $M{=}5$, $\tau{=}0.7$, $R_{\mathrm{K}}{=}R_{\mathrm{V}}{=}2$. Teacher `qwen-vl-max`, probe `qwen-vl-plus`, auditor/judge `qwen-max`, gate `qwen-plus`. Temperatures: 0.0 (gate/audit), 0.2 (VDK), 0.1 (Easy), 0.35 (Hard-K), 0.45/0.75 (Hard-V first/retry). Hard-V neighbor embeddings: Patho-CLIP-L `ViT-L-14`. SFT: full LM update, frozen ViT and projector, AdamW fused (1×10^{-5}, cosine, warmup 0.03), `image_max_pixels`$=262144$, cutoff 4096, seed 42, 3 epochs per stage, DeepSpeed ZeRO-3, `bf16`. Hardware: $7\times$A100-80GB for PathoDC-3B and 7B-no-pretrain; $8\times$A100-80GB for the 7B Quilt-pretrain ablation.

G Question Quality

Admission rates in Table 5 operationalize the three tracks but cannot serve as independent quality scores (Hard-K/Hard-V 100-point cue drops are the filter definitions). We therefore re-evaluate a held-out probe of $n{=}40$ items per track (seed$=$42) with a weaker answerer (`qwen-vl-plus`; Table 6). Gray-image ablation—never used during construction—drops accuracy by **62.5 / 67.5 /**

Structured VDK Generation

```
SYSTEM:
You are a pathology VDK generator. Use the image, caption, and gated
disease entity to produce structured diagnostic knowledge.

The output disease may be:
* identical to the gated disease;
* its canonical medical name;
* a recognized synonym or abbreviation.

It must not represent a different disease. If the two names cannot be
mapped to the same canonical entity, set "status" to "ENTITY_CONFLICT".

MODE may be FULL or REUSE.

In FULL mode:
* generate a concise professional disease description;
* generate 1--2 diagnostically relevant mimics and their distinctions;
* generate key diagnostic criteria;
* identify the anatomical site, histological subtype, and image/stain
  modality when supported;
* generate the present-image diagnosis, visible evidence, and a concise
  explanation linking the evidence to the criteria.

In REUSE mode:
* treat VERIFIED_DESCRIPTION and VERIFIED_CRITERIA as immutable;
* do not rewrite, extend, or contradict them;
* generate only sample-specific differential distinctions, diagnosis,
  visual evidence, and image-specific rationale.

Do not:
* invent patient identifiers, laboratory values, molecular findings,
  clinical history, or image features that are not supported;
* use phrases such as "according to the caption" or "the label says";
* copy the caption as the rationale;
* treat technical identifiers as diagnoses.

Return:
{
  "status": "OK" or "ENTITY_CONFLICT" or "INSUFFICIENT_EVIDENCE",
  "disease_surface": "...",
  "canonical_disease": "...",
  "aliases": ["..."],
  "anatomical_site": "..." or "UNKNOWN",
  "histological_subtype": "..." or "UNKNOWN",
  "modality_or_stain": "..." or "UNKNOWN",
  "description": "...",
  "differentials": [{"disease": "...", "distinction": "..."}],
  "diagnostic_criteria": ["..."],
  "figure_diagnosis": "...",
  "visual_evidence": ["..."],
  "figure_rationale": "..."
}

USER:
MODE: [FULL_OR_REUSE]
Gated disease: [GATED_DISEASE]
Caption: [CAPTION]
Image: [IMAGE]

VERIFIED_DESCRIPTION:
[DESCRIPTION_OR_EMPTY]

VERIFIED_CRITERIA:
[CRITERIA_OR_EMPTY]
```

Figure 5: Prompt template for structured VDK generation (FULL / REUSE modes).

45.0 points on Easy / Hard-K / Hard-V, indicating that questions are image-grounded VQA rather than text artifacts (Hard-K gray still includes k, so the drop is not from reading a knowledge note). Removing k from Hard-K lowers accuracy from 95.0% to 72.5% (Δ_k=22.5); the direction matches the Hard-K claim, but many items remain solvable from the image alone, so we do not claim that Hard-K is unanswerable without k. Replaying the original J_3 protocol on the *same* 40 Hard-V items yields 82.5% dual-image vs. 27.5% single-image accuracy (Δ_{aux}=**55.0**). These independent drops are smaller than the tautological 100-point admission filters, as expected.

Automatic flags on the full Stage II pool (N=109,767; Table 7) complement the probe and are *not* a pathologist audit. Trajectory leak is 0% on every track. Heuristic clinical-like phrasing covers $88.9/93.3/93.4\%$ of Easy/Hard-K/Hard-V, while Hard-K is 73.1% yes/no, matching its criteria-

Mandatory Disease-Level Text Audit

```
SYSTEM:
You are an independent pathology knowledge auditor. Audit the proposed
disease-level knowledge using the gated entity, caption, contextual key,
and generated fields. Do not inspect or make claims about the image.

Accept the record only if ALL conditions hold:

1. the gated name, generated name, canonical name, and aliases refer to
   the same disease entity;
2. the description and diagnostic criteria are professionally coherent
   and do not contradict one another;
3. the knowledge does not improperly mix different anatomical sites,
   histological subtypes, or staining modalities;
4. no generated statement clearly conflicts with the supplied caption;
5. no unsupported patient information, laboratory value, identifier,
   or meta-reference is present;
6. the language is concise, medically appropriate, and structurally
   complete.

A synonym, abbreviation, or canonical name is acceptable when it refers
to the same entity as the gated disease. Reject rather than silently
correct a record containing a disease-identity conflict.

Return:
{
  "decision": "PASS" or "REJECT",
  "entity_consistent": true or false,
  "knowledge_coherent": true or false,
  "context_compatible": true or false,
  "caption_compatible": true or false,
  "language_valid": true or false,
  "failed_criteria": ["..."],
  "brief_justification": "..."
}

USER:
Gated disease: [GATED_DISEASE]
Caption evidence: [EVIDENCE_SPAN]
Caption: [CAPTION]
Contextual key: [CONTEXTUAL_KEY]
Candidate aliases: [ALIASES]
Disease description: [DESCRIPTION]
Diagnostic criteria: [CRITERIA]
```

Figure 6: Prompt template for the mandatory disease-level text audit.

Table 5: Admission under two difficulty levels. J_1: image-only probe; J_2: probe with k; J_3: single- vs. auxiliary-image probe on the rewritten query Q''. Hard-V $J_1{=}0$ is the single-image side of J_3 on Q'', not the Phase 1 test used for Easy and Hard-K. $\Delta_k{=}100$ and $\Delta_{\text{aux}}{=}100$ restate the admission rule and are not independent quality estimates. Easy $\Delta_k{=}-1.4$ shows that, on this filter, Easy items remain solvable without k.

Subset	n	J_1	J_2	J_3 (s/d)	Extra-cue drop†
Easy	90,151	100.0	98.6	–	$\Delta_k{=}-1.4$
Hard-K	13,912	0.0	100.0	–	$\Delta_k{=}100.0$
Hard-V	5,704	0.0‡	–	0.0 / 100.0	$\Delta_{\text{aux}}{=}100.0$

†Full-corpus filter statistic. ‡Single-image side of J_3 on Q''. Held-out quality probe: Table 6.

heavy questions (Figure 9b). Caption 5-gram overlap remains 2.3/3.2% on Easy/Hard-K but 31.0% on Hard-V; disease names occur in 19.6/33.7/70.9% of question strings. The Hard-V rates are inflated by single-field diagnostic wording rather than answer copying—trajectory leak stays 0. Figure 9a lists all 19 site labels: lung, colon, and breast dominate, Other is the unspecified-site bucket, and a named tail of 15 organs (soft tissue, uterus, kidney, ovary, skin, liver, lymph node, bladder, stomach, thyroid, bone, pancreas, brain, prostate, cervix) each remains below 4% per track, with 5,249/3,132/752 distinct disease strings; Easy aspects are balanced, Hard-K is criteria-dominated (75.6%), and Hard-V splits into intra-class Type-A (54.4%) vs. inter-class Type-B (45.6%) contrast. A dual human audit protocol (Appendix H) scores $n{=}50$ items per track on five binary criteria (H_1 visual sufficiency, H_2 gold correctness, H_3 caption independence, H_4 track fidelity, H_5 clinical style).

Image-Level Visual Grounding Audit

```
SYSTEM:
You are an independent pathology visual-grounding auditor. Determine
whether the generated diagnosis and rationale are supported by the
supplied image. Judge the image directly; do not accept a claim merely
because it appears in the caption or disease label.

Accept the record only if ALL conditions hold:

1. each claimed morphological feature is visible or reasonably
   supported by the image;
2. the visible evidence is relevant to the proposed diagnosis;
3. the rationale correctly links image evidence to diagnostic criteria;
4. the rationale does not depend solely on the caption-provided label;
5. no absent tissue structure, cell morphology, stain pattern, or
   clinical information is hallucinated;
6. the evidence is sufficiently specific to be diagnostically useful.

Return:
{
  "decision": "PASS" or "REJECT",
  "supported_features": ["..."],
  "unsupported_claims": ["..."],
  "diagnosis_supported": true or false,
  "rationale_grounded": true or false,
  "caption_label_dependence": true or false,
  "brief_justification": "..."
}

USER:
Image: [IMAGE]
Caption: [CAPTION]
Canonical disease: [CANONICAL_DISEASE]
Diagnostic criteria: [CRITERIA]
Proposed figure diagnosis: [FIGURE_DIAGNOSIS]
Proposed visual evidence: [VISUAL_EVIDENCE]
Proposed figure rationale: [FIGURE_RATIONALE]
```

Figure 7: Prompt template for the image-level visual grounding audit.

Table 6: Independent question-quality probe (n=40 per track, seed=42; answerer `qwen-vl-plus`). Gray-image ablation was *not* used during construction. $\Delta_{\text{img}} = \text{Acc}_{\text{full}} - \text{Acc}_{\text{gray}}$ (percentage points). Cue ablation tests the extra evidence each track claims to need: Hard-K removes k; Hard-V removes V_{aux} under the original J_3 judge. Easy has no extra cue, so Δ_{cue} is undefined.

	Image grounding			Track-specific extra cue		
Track	Full	Gray	$\Delta_{\text{img}}\uparrow$	With	w/o	$\Delta_{\text{cue}}\uparrow$
Easy	72.5	10.0	62.5	–	–	–
Hard-K	95.0	27.5	67.5	95.0	72.5	22.5
Hard-V	77.5†	32.5†	45.0†	82.5‡	27.5‡	**55.0**‡

†Loose probe (gray-image was never a construction filter). ‡Same 40 Hard-V items, original J_3 (`run_j3_median` + `qwen-max`); not commensurate with 77.5.

H Human Audit of Constructed VQA

This appendix specifies the dual human audit used as an independent check of question quality after J_1/J_2/J_3. The audit is *not* an admission filter (Sec. 3.3). Adjudicated labels accompany the data release.

H.1 Sampling

We draw a stratified sample of n=50 items per track (Easy / Hard-K / Hard-V), N=150 in total, uniformly without replacement from `train_easy.jsonl` / `train_mid_all.jsonl` / `train_hard_all.jsonl` with seed 42. The 40 independent-probe qids per track (Table 6) are included when present so that human judgments can be aligned with the automatic gray-image and extra-cue ablations; the remaining 10 per track are fresh draws. Annotators do not see automatic leak flags, VL predictions, or PathoDC-Bench labels.

Question-Conditioned Leakage Filter

```
SYSTEM:
You are a diagnostic-knowledge leakage auditor. Determine whether the
knowledge paragraph allows the reference answer to be copied directly
without interpreting the query image.

Check for:

1. an exact occurrence of the answer;
2. canonical names, abbreviations, or aliases of the answer;
3. an explicit semantic paraphrase that directly states the answer;
4. a figure-specific diagnosis or conclusion;
5. copied label-bearing phrases from the source caption;
6. meta-cues such as "according to the caption" or "the diagnosed case";
7. a target disease name that was not replaced by [TARGET DISEASE].

Do not treat useful diagnostic criteria as leakage merely because they
help answer the question when combined with the image. The distinction
is whether the answer can be recovered by direct textual lookup rather
than image--knowledge reasoning.

When leakage is localized, remove or mask only the offending span while
preserving the diagnostic criteria. Reject the sample if leakage cannot
be removed without destroying the knowledge's diagnostic utility.

Return:
{
  "decision": "PASS" or "SANITIZE" or "REJECT",
  "leakage_detected": true or false,
  "leakage_types": ["..."],
  "leaked_spans": ["..."],
  "sanitized_knowledge": "...",
  "answer_copyable_without_image": true or false,
  "brief_justification": "..."
}

USER:
Question: [QUESTION]
Reference answer: [ANSWER]
Answer aliases: [ANSWER_ALIASES]
Source caption: [CAPTION]
Delexicalized knowledge: [KNOWLEDGE]
```

Figure 8: Prompt template for the question-conditioned leakage filter.

Table 7: Automatic question-hygiene flags on the full Stage II pool (N=109,767). Not a human audit. Clin.: heuristic clinical-like lexicon. Y/N: `yes_no` questions (Hard-K matches the criteria track). Cap.: question looks like a caption. Traj.: gold answer appears in the rationale. n-gram: question shares a 5-gram with the source caption. Dx: question string contains the disease name (inflated on Hard-V by single-field diagnostic wording).

Track	n	Clin.	Y/N	Cap.	Traj.	n-gram	Dx
Easy	90,151	88.9	9.0	7.3	0.0	2.3	19.6
Hard-K	13,912	93.3	73.1	7.6	0.0	3.2	33.7
Hard-V	5,704	93.4	2.8	5.2	0.0	31.0	70.9

H.2 Interface and allowed evidence

Each item is shown as the query image V, question Q, and gold answer A. Track-specific extra evidence is included because H_1 and H_4 are defined on the same evidence pack as construction:

- Easy: (V, Q, A) only.
- Hard-K: (V, Q, A, k) with the leakage-safe prior k.
- Hard-V: $(V, Q, A, V_{\text{aux}})$ with the auxiliary neighbor, plus the neighbor class (Type-A same-disease vs. Type-B confusable).

The source caption is available on a second screen *only* for H_3 (caption independence) and is hidden while scoring H_1, H_2, H_4, and H_5. The intended track name is visible because the evidence pack depends on it.

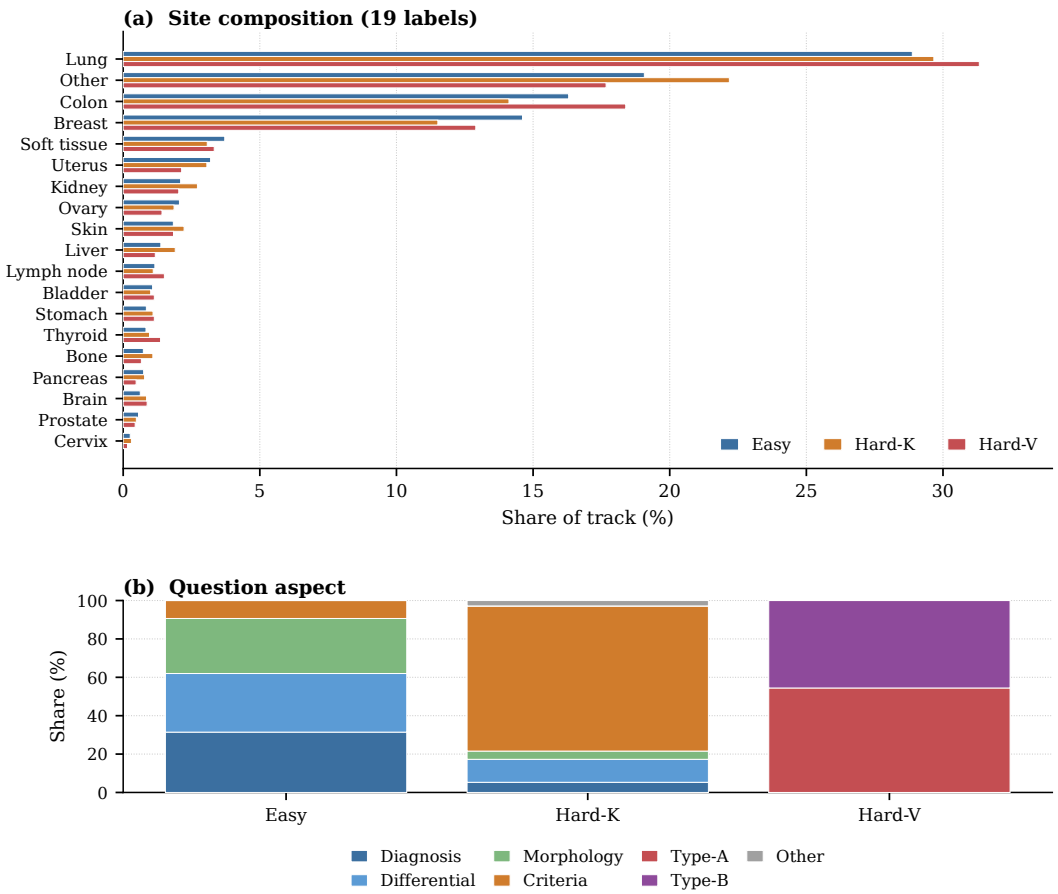


Figure 9: Site and aspect diversity of the constructed Stage II pool (N=109,767). (a) Share of each track for all 19 organ labels, sorted by overall frequency. Other is the unspecified-site label. The remaining 15 named sites are soft tissue (3.6%), uterus (3.1%), kidney (2.2%), ovary (2.0%), skin (1.9%), liver (1.4%), lymph node (1.2%), bladder (1.1%), stomach (0.9%), thyroid (0.9%), bone (0.8%), pancreas (0.7%), brain (0.7%), prostate (0.5%), and cervix (0.3%) of the full pool. (b) Question aspect. Easy is balanced across diagnosis, differential, and morphology. Hard-K is criteria-heavy (75.6%). Hard-V items are contrastive by construction: Type-A same-disease (54.4%) vs. Type-B confusable (45.6%). Distinct disease strings: 5,249 / 3,132 / 752.

H.3 Five binary criteria

Every rater assigns $\{0, 1\}$ independently on five questions. 1 means pass.

H_1 **Visual sufficiency.** Given the track's allowed evidence (not the caption), can a trained reader answer Q without guessing? Score 1 if the evidence supports a definite answer; 0 if the image is unrelated, blank, too cropped, or the question is unanswerable from the given inputs. For Hard-K, k may be used; for Hard-V, V_{aux} may be used. This is *not* "Easy questions must be solvable from V alone"—that belongs to H_4.

H_2 **Gold correctness.** Is A a valid answer to Q under the same evidence as H_1? Score 1 if A is correct or an acceptable clinical synonym; 0 if A contradicts the image, answers a different question, or is an empty/placeholder string.

H_3 **Caption independence.** After revealing the source caption: are Q and A more than a near-paraphrase of that caption? Score 1 (no leak) if answering still requires looking at the slide (and k

or V_{aux} when allowed); 0 if Q copies a caption clause or A is spanned verbatim by the caption so that the item is text QA.

H_4 **Track fidelity.** Does the item need the evidence that its track claims?

- Easy: 1 if (V, Q) suffices and k/V_{aux} are unnecessary.
- Hard-K: 1 if (V, Q) is insufficient or highly ambiguous, and k is what makes A recoverable.
- Hard-V: 1 if a single image is insufficient and comparing V with V_{aux} is what resolves A.

Score 0 on Easy items that secretly need external knowledge, Hard-K items that are solvable from the image alone with no use of k, or Hard-V items that are solvable without the neighbor.

H_5 **Clinical style.** Is Q a diagnostic/pathology question (criteria, morphology, differential, diagnosis) rather than caption completion ("what does the image show in words")? Score 1 for clinical wording; 0 for generic caption-like prompts.

H.4 Aggregation

Two annotators with pathology training score every item independently. A third annotator adjudicates disagreements (any criterion where the two labels differ). Reported pass rates use the *adjudicated* label. Cohen's κ is computed on the two primary raters *before* adjudication, separately for each H_i. Item-level Overall is $H_1 \wedge H_2 \wedge H_3$ after adjudication; H_4 and H_5 are reported but not required for Overall, because a well-formed VQA item can still be off-track or slightly caption-like. Track-level pass rate is the mean of item Overall in that track. We do not use 1–10 impression scores.

I PathMMU

Evaluation setting. PathMMU (Sun et al., 2025b) measures transfer to expert-validated atlas/textbook questions rather than alignment with the PathoDC evidence taxonomy. We report multiple-choice accuracy on Atlas, EduContent, PathCLS, and PubMed, and use `qwen-vl-max` only for failed option parses. Table 8 reports test-tiny and test, and Table 9 reports val. PathoDC-3B and PathoDC-7B are the sequential Easy→Hard-K→Hard-V finals.

Results. PathoDC-3B gains 10.73 points over Instruct on test All (57.43) and narrowly exceeds Patho-R1-3B (56.27). PathoDC-7B gains 18.22/15.50 points on test-tiny/test All and leads the compared 7B/8B models. PathGen-LLaVA-13B remains strongest on test (67.89), and the advantage is not uniform on val, bounding the transfer claim.

Table 8: PathMMU multiple-choice accuracy (%). Left: test-tiny; right: test. Subsets: Atlas, EduContent (Edu.), PathCLS (CLS), and PubMed. PathoDC-3B / PathoDC-7B: sequential Easy→Hard-K→Hard-V finals. Best / second-best *within 3B and within 7B/8B*: **bold** / underlined. PathGen-LLaVA-13B is listed separately and is not used for those marks.

	Model	PathMMU test-tiny					PathMMU test				
		Atlas	Edu.	CLS	PubMed	All	Atlas	Edu.	CLS	PubMed	All
General VLM	Qwen2.5-VL-3B	57.21	52.55	29.94	55.16	50.05	53.44	50.15	27.63	53.86	46.70
Pathology VLM	Patho-R1-3B	**75.00**	**64.31**	**38.42**	66.19	**62.32**	71.34	60.43	35.05	61.86	56.27
	PathoDC-3B	74.40	63.87	37.80	**67.25**	62.27	**72.32**	**61.34**	**36.23**	**63.21**	**57.43**
General VLM	InternVL2-8B	50.00	51.76	34.46	60.14	50.60	48.81	49.20	32.23	50.81	45.79
	InternVL2.5-8B	59.62	61.18	35.59	60.14	55.59	59.82	54.43	34.44	56.19	51.04
	InternVL3-8B	61.54	62.35	44.63	64.41	59.39	60.45	59.00	44.24	60.93	56.46
	Qwen2.5-VL-7B	53.37	56.47	31.07	55.16	50.49	54.57	51.16	29.60	53.50	47.40
Pathology VLM (13B)	PathGen-LLaVA-13B	81.25	71.37	58.19	75.09	72.20	76.22	69.34	59.07	69.79	67.89
Pathology VLM	Quilt-LLaVA-v1.5-7B	41.35	45.10	26.55	36.30	38.00	36.55	37.97	25.12	33.80	33.08
	HuatuoGPT-Vision-7B	64.90	59.61	35.59	65.48	57.98	59.45	56.21	34.25	61.39	53.49
	Patho-R1-7B	80.30	**69.00**	46.30	71.20	67.86	75.70	65.20	46.20	65.00	61.84
	PathoDC-7B	**81.30**	68.74	**47.32**	**72.85**	**68.71**	**76.30**	**66.70**	**46.57**	**66.32**	**62.90**

Table 9: PathMMU-val multiple-choice accuracy (%). PathoDC-3B / PathoDC-7B: sequential Easy→Hard-K→Hard-V finals. Best / second-best: **bold** / underlined.

	Model	Atlas↑	Edu.↑	CLS↑	PubMed↑	All↑
General VLM	Qwen2.5-VL-3B	57.5	47.3	29.2	50.6	47.0
Pathology VLM	Patho-R1-3B	**73.8**	**60.3**	32.3	**58.8**	**56.8**
	PathoDC-3B	55.0	37.7	**35.4**	46.8	43.6
General VLM	InternVL2-8B	52.5	45.9	34.4	48.1	45.8
	InternVL2.5-8B	61.3	51.4	34.4	57.9	52.6
	InternVL3-8B	57.5	53.4	42.7	58.4	54.2
	Qwen2.5-VL-7B	63.7	47.9	29.2	51.9	48.6
Pathology VLM	PathGen-LLaVA-13B	80.0	**69.2**	**57.3**	**69.5**	**68.8**
	Quilt-LLaVA-v1.5-7B	41.2	40.4	27.1	37.8	37.1
	HuatuoGPT-Vision-7B	66.2	54.8	31.2	58.4	53.9
	Patho-R1-7B	**81.2**	64.4	49.0	66.5	65.0
	PathoDC-7B	70.0	49.3	32.3	62.7	55.0

J STAGE VARIANTS

The main stage-variant comparison is Table 3 in Sec. 4.4. Figure 10 plots sequential Easy→Hard-K→Hard-V checkpoints against Instruct Base, including the Quilt caption-pretrain valley. Figure 11 shows the matched-split Bench bars of those checkpoints; together, the two figures visualize the stage-wise transfer and matched-split specialization summarized numerically in Table 3.

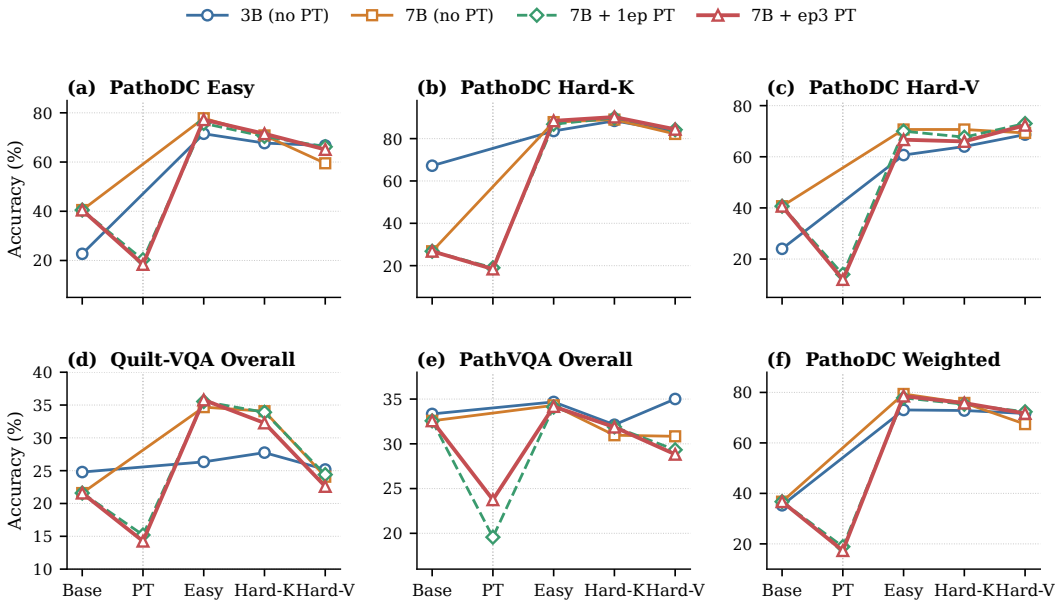


Figure 10: Stage-wise accuracy (%) along the curriculum. The x-axis is Base (Instruct), Quilt caption pretrain (PT; 1ep/ep3 only), Easy SFT, Easy→Hard-K, and Easy→Hard-K→Hard-V. 3B/7B (no PT) skip the PT tick. (a–c) PathoDC-Bench splits: each constructed stage peaks its matched test split. (d–e) Public Quilt-VQA / PathVQA Overall: Easy SFT transfers off-domain; 7B Hard-V recedes. (f) PathoDC-Bench Weighted. PT is a valley: caption pretrain alone falls below Base; Easy SFT recovers.

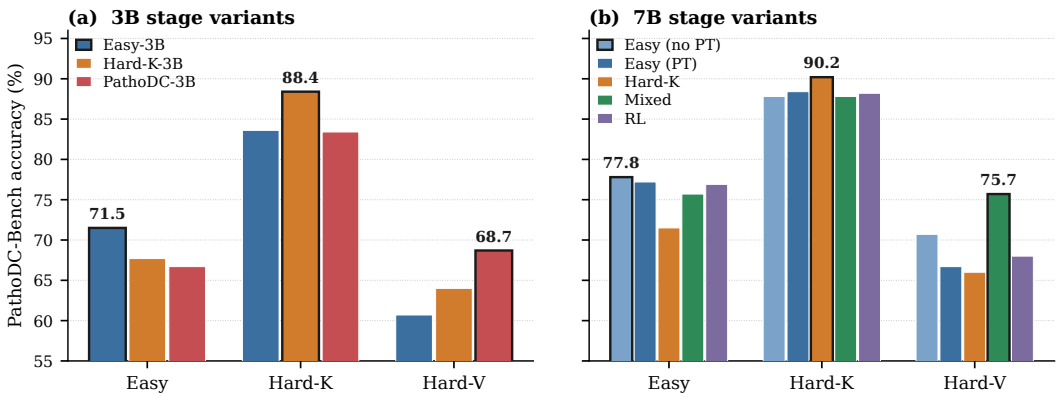


Figure 11: PathoDC-Bench accuracy (%) of the Table 3 stage variants, grouped by *test* split. The winner of each difficulty is the tallest bar in that cluster (black edge, labeled). A line plot versus Easy→Hard-K→Hard-V would be misleading: Hard-K is high for every checkpoint, so the matched peak is column-wise, not a peak along the x-axis. (a) 3B: Easy / Hard-K / PathoDC-3B peak their matched split. (b) 7B: PathoDC-Easy-7B peaks Easy, PathoDC-Hard-K-7B peaks Hard-K, and PathoDC-7B is the sequential Hard-V final. Caption-pretrain-only is omitted (18.4 / 18.4 / 12.0). y-axis starts at 55 to separate the competing checkpoints.